Dimension Reduction in the Atmospheric Sciences

Imola K. Fodor and Chandrika Kamath
Center for Applied Scientific Computing
Lawrence Livermore National Laboratory
fodor1@llnl.gov, kamath2@llnl.gov

November 29, 2002

Abstract

Improved data collection capabilities and increased computational power have led to global observations being collected at an ever increasing pace, and global simulation models generated at an ever increasing resolution. As a result, the atmospheric sciences are a rich source of high-dimensional datasets. In this paper, we describe the use of principal component analysis and independent component analysis as two ways of reducing the dimension of climate data sets. We show how these techniques can be used to separate effects of volcanic eruptions and ENSO variations from global temperatures.

1 Introduction

Understanding changes in the global climate is a challenging scientific problem. Research groups around the world have been developing different models in order to characterize accurately the different processes involved and their interactions. With increased computational resources, the complexities of the resulting models and the quantity of the output data have also increased dramatically. To advance our scientific understanding, we need data analysis techniques that can keep up with this information flow.

One of the problems in the atmospheric sciences is to identify the main sources of variability in the climate and to decompose the data into the contributions of the different sources. Such a decomposition would enable scientists to compare directly the predictions based on different simulation models and also assess the effect of human contributions on the global climate. For example, volcano eruptions and El Niño and Southern Oscillation (ENSO) variations both influence global temperatures. Recent eruptions of the El Chicón and Mt. Pinatubo volcanoes temporally coincided with large ENSO events, which makes the separation of the volcano effect from the ENSO effect difficult [5].

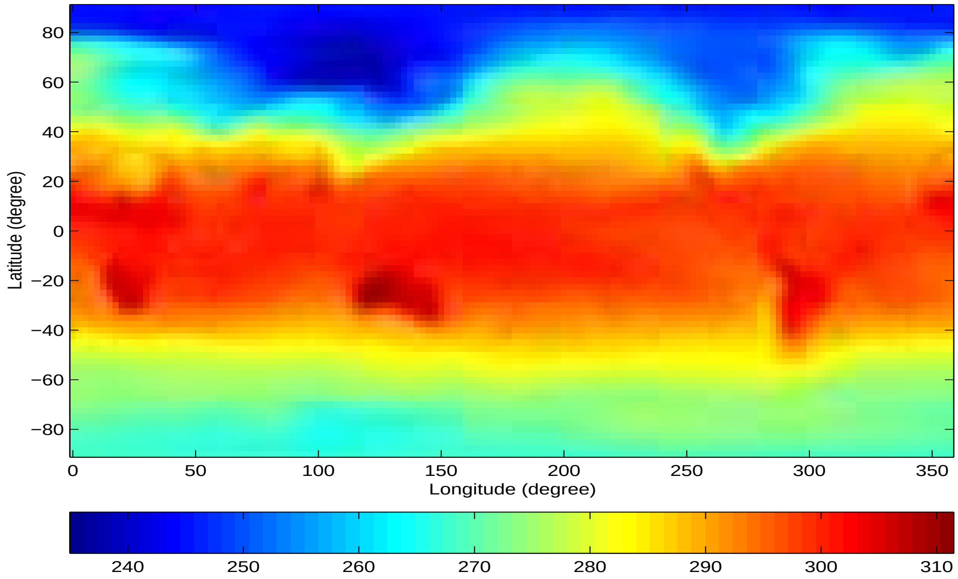


Figure 1: January 1979 raw temperatures (Kelvin) on the 144 longitudes by 73 latitudes spatial grid for pressure level 1000 hPa.

Principal component analysis (PCA) is a dimension-reduction method that can be applied in various contexts in the atmospheric sciences [4]. Given a high-dimensional data set, PCA seeks orthogonal linear projections in a lower dimensional space with largest variance. Independent component analysis (ICA) [2] is a related technique that seeks independent linear projections.

We investigate PCA and ICA as two techniques to reduce the dimension of global temperature series and to identify the main sources of variability. Using reanalysis data from computer simulations, we illustrate these methods and demonstrate their usefulness.

2 Description of the Data

We conducted this study using monthly mean temperature reanalysis data [3] from the National Centers for Environmental Prediction (NCEP). The data is on a 144×73 longitude-by-latitude grid, on 17 vertical pressure levels ranging from 10000 hPa close to the surface of the earth to 10 hPa at the highest elevation. It spans 264 months, from January 1979 to December 2000. Figure 1 displays the raw temperatures for January 1979 on the 144×73 latitude by longitude grid at the 1000 hPa pressure level.

Since we expected the ENSO and volcano signals to have a strong latitudinal dependence, we performed our analyses on zonally averaged data. At a given month and level, we first calculated the 73 zonal means over the 144 longitudes. Next, following standard practices in the atmospheric sciences [4], we removed the seasonal variation and centered the data as follows. For each month, we replaced the values at each of the $73 \times 17 (= 1241)$ latitude-by-level grid points by subtracting their corresponding monthly means over the entire 22-year period. The resulting time-centered values are referred to as anomalies in the atmospheric sciences literature.

Next, we performed PCA and ICA on an $n = 264$ by $p = 1241$ time-by-space dimension anomaly data set $\mathbf{Z}_{n\times p}$. This data set was first weighted appropriately to account for the unequal spatial grid sizes and altitude bands as shown in Figure 2. The latitude weight in Figure 2 (a) is largest at the equator and decreases smoothly to zero toward the poles. The pressure level weights in Figure 2 (b) account for the unequal altitude bands in the data. For example, the width of the fourth altitude band is largest, so that its corresponding weight is also the largest. The combined weights in Figure 2 (c) are the products of the corresponding latitude and level weights. Figure 3 displays the January 1979 anomaly temperatures after the weights have been applied to the time-centered zonally averaged data.

3 Analysis Results with PCA

The PCA technique seeks linear combinations of the data with maximal variance. Given the zonal anomaly dataset $\mathbf{Z}_{n\times p}$ of dimension $n = 264$ months at $p = 1241$ spatial grid points, let

$$\mathbf{S}_{p\times p} = \mathbf{Z}'_{p\times n}\mathbf{Z}_{n\times p} \tag{1}$$

denote the state-space scatter matrix [4], with $\mathbf{Z}'$ denoting the transpose of the matrix $\mathbf{Z}$. The eigenvalue decomposition of $\mathbf{S}$ provides the eigenvector matrix $\mathbf{E}_{p\times p}$ and the eigenvalue matrix $\mathbf{D}_{p\times p}$ such that

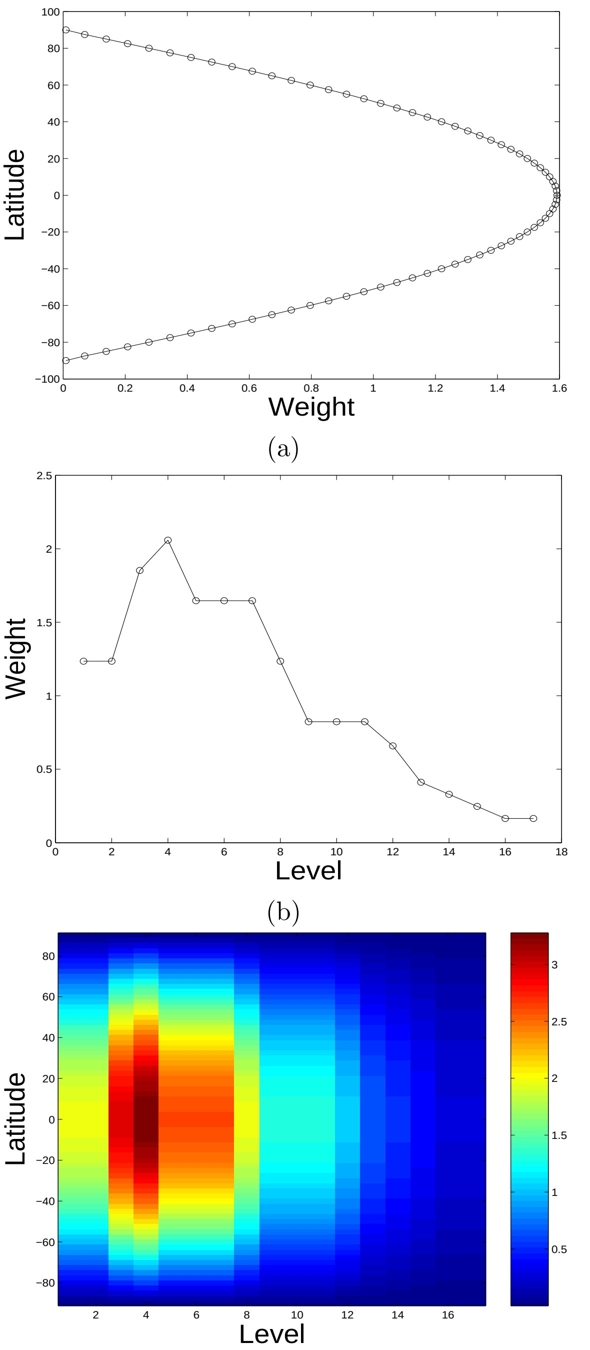
$$\mathbf{S}_{p\times p} = \mathbf{E}_{p\times p}\mathbf{D}_{p\times p}\mathbf{E}'_{p\times p}, \tag{2}$$

where $\mathbf{E}$ is orthonormal, $\mathbf{E}'\mathbf{E} = \mathbf{E}\mathbf{E}' = \mathbf{I}_{p\times p}$, and $\mathbf{D}$ is diagonal, $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_p)$ with $d_1 \geq d_2 \ldots \geq d_p$. To make the eigenvectors unique, we adopt the convention that the first non-zero term in each of the vectors is positive. Only $r = \mathrm{rank}(\mathbf{Z}_{n\times p})$ eigenvectors are well defined. In our case, since $n < p$, and the anomalies at each spatial grid sum to zero over time, $r = n - 1$. The corresponding state-space principal components are

$$\mathbf{A}_{n\times p} = \mathbf{Z}_{n\times p}\mathbf{E}_{p\times p}. \tag{3}$$

Since the eigenvalue decomposition of the $p\times p$ state-space scatter matrix $\mathbf{S}$ in Eq. (2) is computationally expensive, we can apply the following computational shortcut [4] based on the eigenvalue decomposition of the smaller dimensional $n \times n$ sample-space scatter matrix

$$\mathbf{T}_{n\times n} = \mathbf{Z}_{n\times p}\mathbf{Z}'_{p\times n}. \tag{4}$$



(a)

(b)

(c)

Figure 2: Weights for the zonally averaged data. (a) The 73 latitude weights sum to 73. (b) The 17 pressure level weights sum to 17. (c) The 73 × 17 combined latitude by pressure level weights sum to 1241.

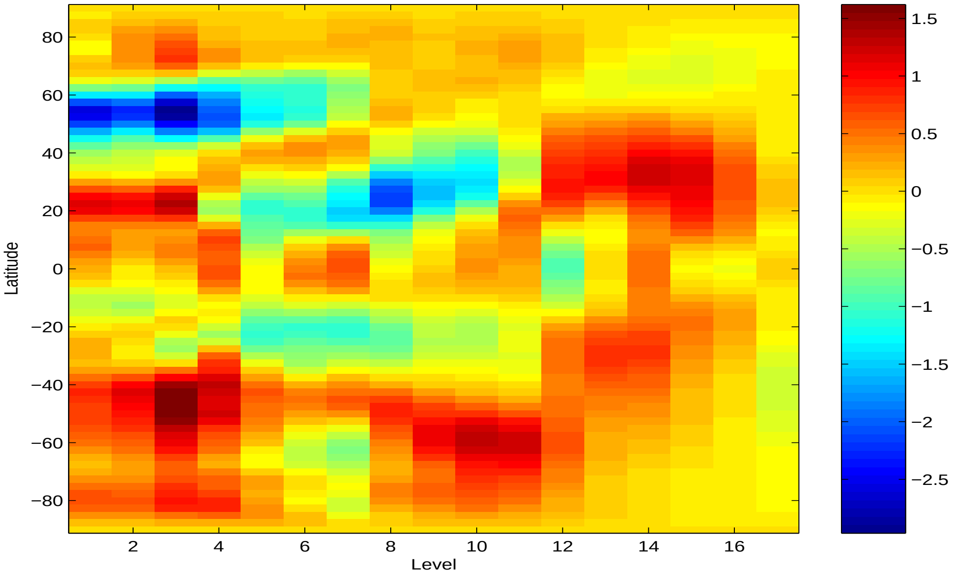


Figure 3: January 1979 zonal anomaly temperatures on the 17 pressure levels.

Let

$$\mathbf{T}_{n\times n} = \mathbf{F}_{n\times n}\mathbf{M}_{n\times n}\mathbf{F}'_{n\times n}, \tag{5}$$

denote its associated eigenvalue decomposition, with $\mathbf{F}$ orthonormal, that is, $\mathbf{F}'\mathbf{F} = \mathbf{F}\mathbf{F}' = \mathbf{I}_{n\times n}$, and $\mathbf{M}$ diagonal, $\mathbf{M} = \text{diag}(m_1, \ldots, m_n)$ with $m_1 \geq m_2 \ldots \geq m_n$. The sample-space principal components are given by

$$\mathbf{B}_{p\times n} = \mathbf{Z}'_{p\times n}\mathbf{F}_{n\times n}. \tag{6}$$

It is easy to show [4] that the first r state-space eigenvectors in $\mathbf{E}$ can be obtained from the sample-space principal components of $\mathbf{B}$ by the following transformation

$$\hat{E}_{i,j} = B_{i,j}/\sqrt{m_j}, \quad i = 1, \ldots, p;\; j = 1, \ldots, r. \tag{7}$$

Figure 4 presents the first six state-space eigenvectors, that is, the first six columns of $\mathbf{E}_{p\times p}$, re-mapped to the 73 × 17 latitude-by-pressure level spatial grid. Figure 5 displays the corresponding first six principal component time series, which are the first six columns of $\mathbf{A}_{n\times p}$.

The patterns in the first eigenvector and corresponding principal component suggest characteristics associated with ENSO variations. To investigate this further, we consider the first state space PC time series (scaled) along with the corresponding Niño 3.4 index obtained from NCEP as shown in Figure 6. The Niño 3.4 index measures the monthly sea surface temperature deviation from its long-term mean averaged over the Niño 3.4 region (5N-5S, 120-170W). It is a standard index used to monitor ENSO events. If its 5-month running average

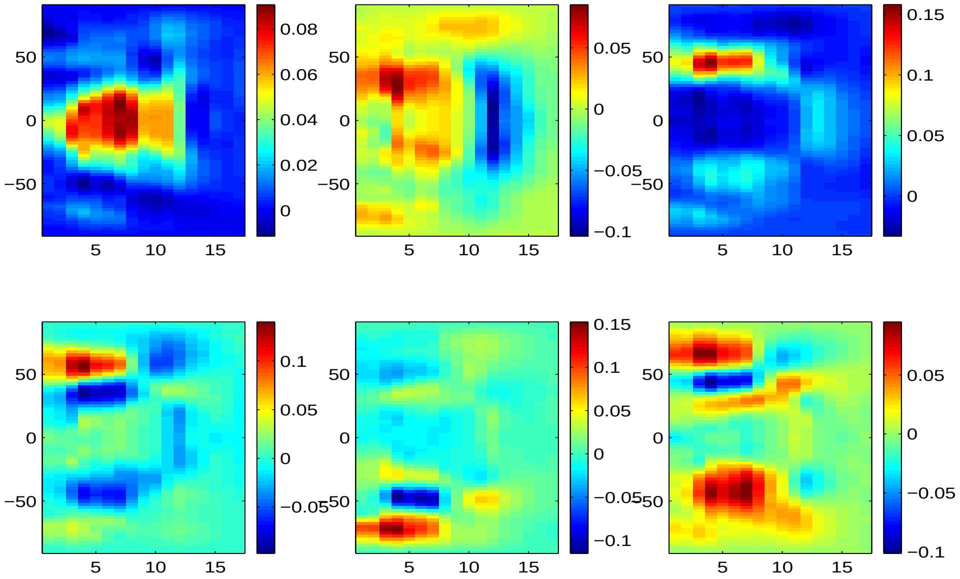


Figure 4: First 6 state space eigenvectors.

exceeds +0.4 (-0.4) degrees Celsius, it is considered an El Niño (La Niña) event. The correlation coefficient of the two series in Figure 6 is equal to 0.4940. The plots indicate that the two series follow a similar pattern, but they are shifted relative to each other. The locations of the three major peaks suggest that we can increase the correlation between the two series by shifting the PC series a few months back. Indeed, lagging the PC series by one month increases the correlation coefficient to $\text{cor}(1) = 0.6059$. The correlation increases monotonically up until the lag is five months, cor(2)=0.6689, cor(3)=0.7074, cor(4)=0.7186, cor(5)=0.7227, then it starts to decline monotonically, cor(6)=0.6954. Figure 7 presents the PC series shifted back 5 months along with the Niño region 3.4 index.

The PCA decomposition can be used as an effective dimension reduction tool. The original anomaly data $\mathbf{Z}_{n\times p}$ can be perfectly reconstructed from the p eigenvectors as

$$\mathbf{Z}_{n\times p} = \mathbf{Z}_{n\times p}\mathbf{E}_{p\times p}\mathbf{E}'_{p\times p}. \tag{8}$$

Considering only the first k eigenvectors, the reconstruction becomes

$$\hat{\mathbf{Z}}^{(k)}_{n\times p} = \mathbf{Z}_{n\times p}\mathbf{E}_{p\times k}\mathbf{E}'_{k\times p}. \tag{9}$$

The kth cumulative sum of the eigenvalues normalized by the total sum of the eigenvalues defined as

$$\lambda(k) = \frac{\sum_{i=1}^{k} d_i}{\sum_{i=1}^{p} d_i}, \tag{10}$$

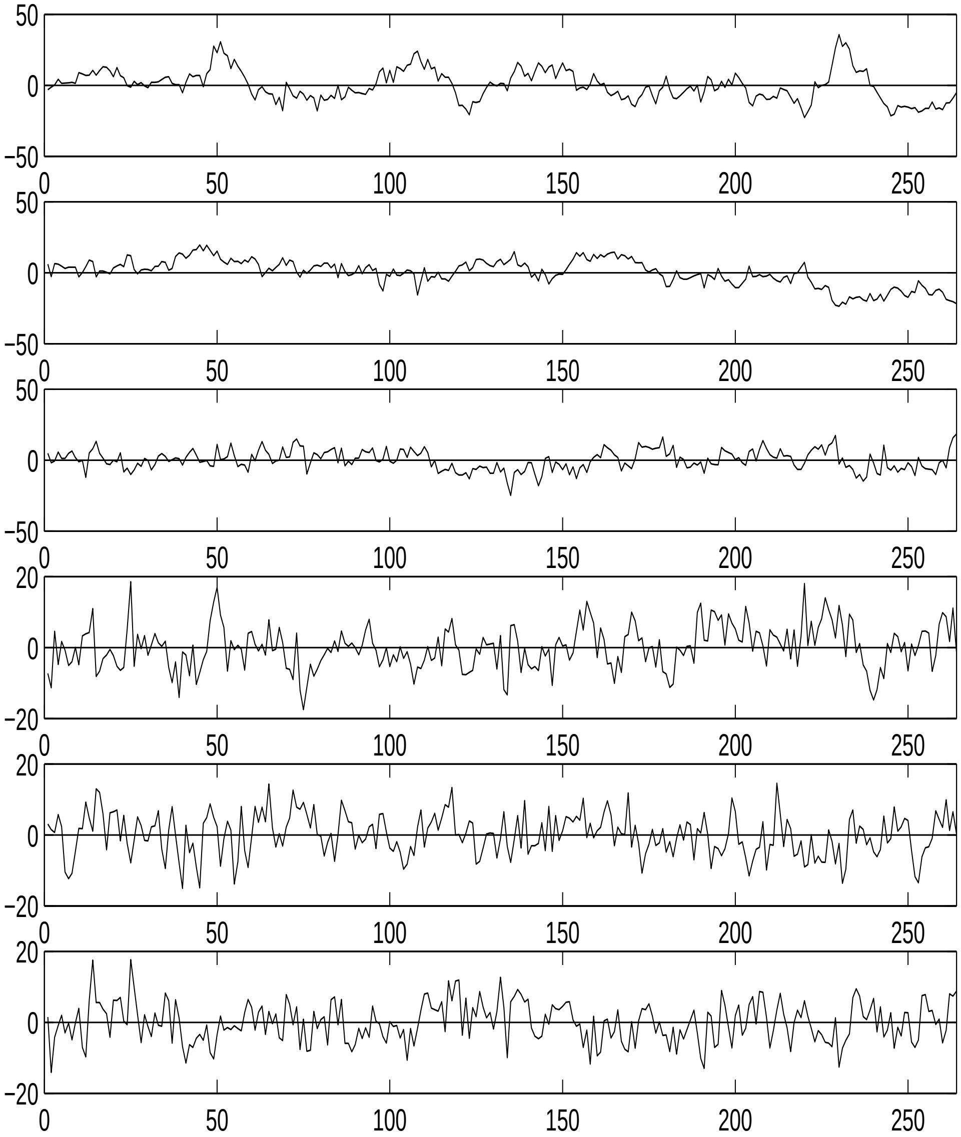


Figure 5: First 6 state space PC time series.

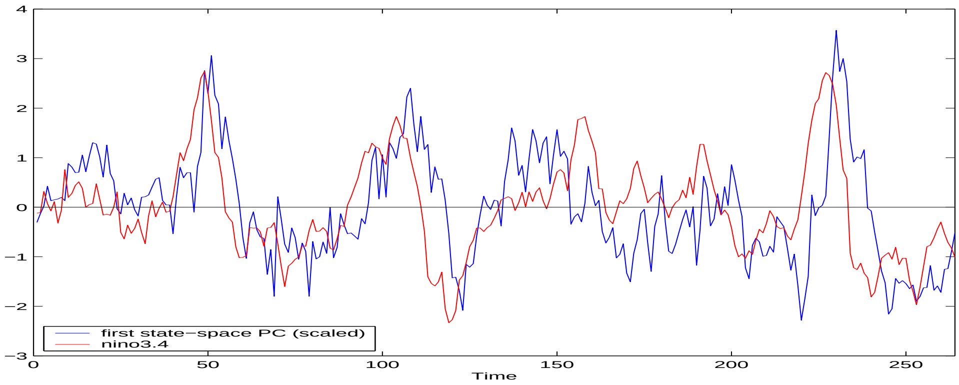


Figure 6: First state space PC time series (scaled), and Niño region 3.4 sea surface temperature anomaly series. Correlation coefficient is 0.49.

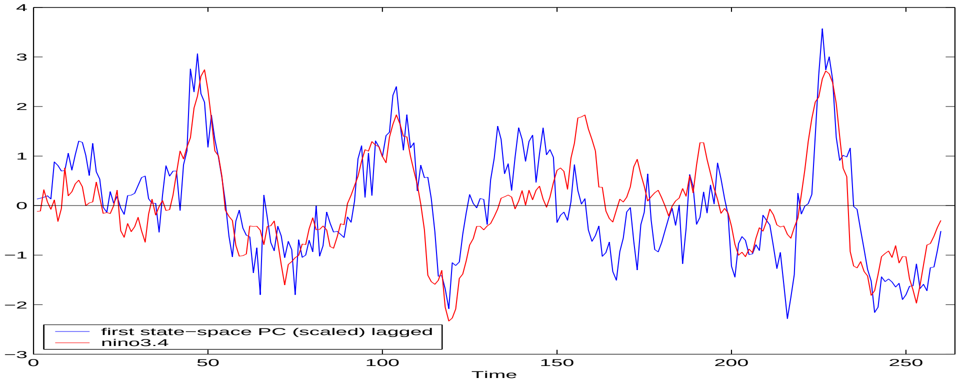


Figure 7: First state space PC time series (scaled) with a five month lag, and Niño region 3.4 sea surface temperature anomaly series. Correlation coefficient is 0.73.

k	$\lambda(k)$	k	$\lambda(k)$
1	0.2033	22	0.9003
2	0.3527	35	0.9501
3	0.4390	50	0.9738
4	0.5057	79	0.9901
5	0.5660	100	0.9946
6	0.6212	150	0.9986
7	0.6626	200	0.9997
8	0.7006	216	0.9999
9	0.7361	234	1.0000
10	0.7645	1241	1.0000

Table 1: Percent of the total variation explained $\lambda(k)$ by first k principal components.

indicates how much of the variation is explained by the first k, $k \leq p$, components. Table 1 shows the percent of variation explained for selected values of k.

Figure 8 (a) shows the original anomaly data for January 1979. Based on the values in Table 1, the first k=5 principal components explain 57% of variance. Figure 8 (b) illustrates the approximation of the anomaly data using $k = 5$ in Eq. (9). The approximation discerns a main cold pattern around the 50 degree latitude line over the first eight or so levels, another less pronounced negative region extending from latitudes -20 to 20 over levels four to ten, and a warm pattern around the same latitude band, extending over pressure levels eleven to fourteen. Some of the overall features in the original data are already apparent in this crude approximation, but many important details are missing. Figure 8 (c) displays the reconstruction obtained by using the first $k = 10$ principal components, accounting for 76% of the total variance in the data. As we increase the number of principal components used in the reconstruction, the quality of the approximation improves. Figures 8 (d), (e) and (c) present the corresponding results for $k = 22$, 35 and 79, accounting for 90%, 95% and 99% of the variation, respectively. As the last panel indicates, using 79 eigenvectors from the total of 1241 provides an excellent representation the data. The information in the original high-dimensional data can be reconstructed without significant loss of accuracy from the lower-dimensional approximation.

4 Analysis Results with ICA

Given a high-dimensional dataset, ICA seeks linear projections that are mutually independent. The simplest ICA model [2] assumes that the observed signal $\mathbf{Z}_{n\times p}$ is a linear mixture of the unobservable independent sources $\mathbf{C}_{n\times p}$, with

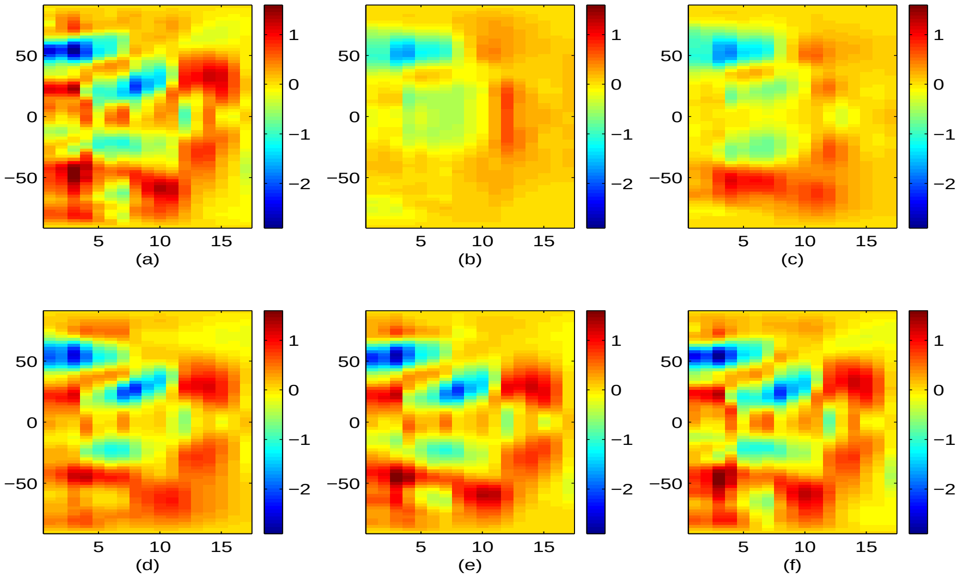


Figure 8: Successive PC approximations. (a) Original January 1979 anomaly data. (b) Approximation based on the first k=5 PCs. (c) Approximation based on the first k=10 PCs. (d) Approximation based on the first k=22 PCs. (e) Approximation based on the first k=35 PCs. (f) Approximation based on the first k=79 PCs.

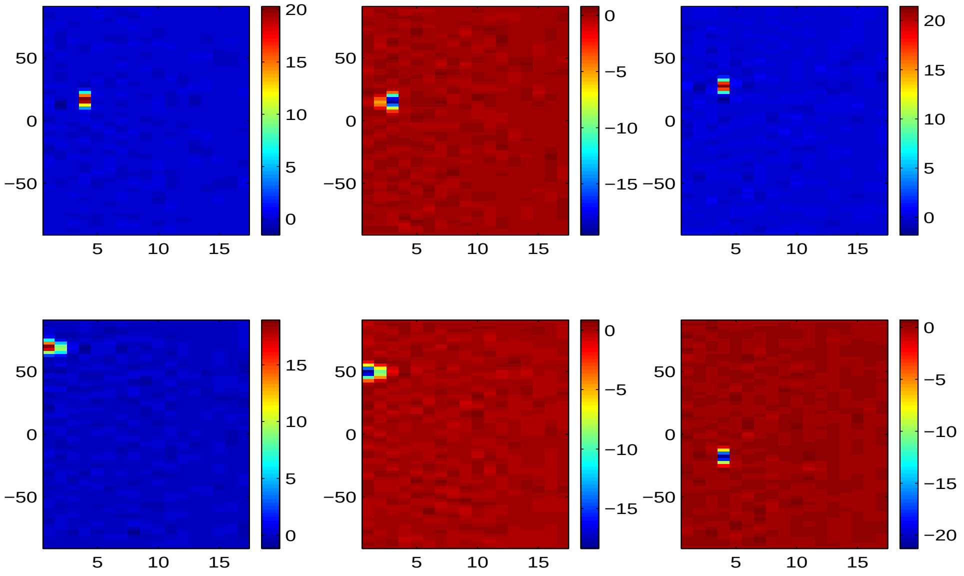


Figure 9: Six independent component bases obtained from the full-dimensional anomaly data.

the mixing matrix $\mathbf{M}_{n\times n}$,

$$\mathbf{Z}_{n\times p} = \mathbf{M}_{n\times n}\ \mathbf{C}_{n\times p}. \tag{11}$$

ICA methods use the observations to estimate the mixing matrix and the independent components by first specifying a measure of independence, and then providing an algorithm to optimize that measure of independence. There are several different variants of ICA, depending on the independence measure and optimization method. Example include the FastICA and the maximum likelihood based gradient ascent approaches [2].

PCA can be considered a special case of ICA where the optimization criteria maximizes the variance, with the constraint that the resulting estimates be orthonormal. Note that for Gaussian signals, orthogonality is equivalent to independence, and ICA only makes sense for non-Gaussian signals.

If the physical processes contributing to the overall climate are non-Gaussian, independent, mixed in a linear fashion, then we can expect simple ICA methods to perform better than the PCA in separating the underlying sources.

Figure 9 shows the first six independent component estimates with the FastICA method applied to the $\mathbf{Z}_{n\times p}$ anomaly data. The resulting IC estimates are very localized spatially and are not interpretable scientifically in terms of atmospheric processes. The large spikes are in fact artifacts due to over-learning in the ICA estimation [2]. Estimating n independent components, along with the mixing matrix as well, from only n samples is statistically impossible since there are more parameters to estimate than available observations. We do have n samples of the p-dimensional signal, but we do not expect to have n truly independent component sources.

Instead of working with the full high-dimensional anomaly data, we explore ICA on reduced dimensional representations of the data. Since there are only a limited number of atmospheric processes, before applying ICA we first need to reduce the data to k dimensions, where k is the number of independent components sought. The PCA results in Section 3 indicate that the first $k = 22$ PCs explain 90% of the variance. So, we applied ICA to the anomaly data of reduced dimension, as described by the first $k = 22$ principal components. Since the original anomalies are linear combinations of the PCs, we can easily obtain the IC coefficients corresponding to the original data from the IC coefficients obtained for the PCs.

Let the rows of $\mathbf{C}_{k\times p}$ denote the k independent component estimates obtained by applying ICA to the first $k = 22$ eigenvectors of the scatter matrix $\mathbf{S}_{p\times p}$ in Eq. (2)

$$\mathbf{E}'_{k\times p} = \mathbf{M}_{k\times k}\ \mathbf{C}_{k\times p}. \tag{12}$$

Combining the results in Eq. (12) with the decomposition in Eq. (9), the original anomaly data can be written in terms of the estimated independent components as

$$\hat{\mathbf{Z}}^{(k)}_{n\times p} = \mathbf{Z}_{n\times p}\ \mathbf{E}_{p\times k}\ \mathbf{M}_{k\times k}\ \mathbf{C}_{k\times p} = \mathbf{N}_{n\times k}\ \mathbf{C}_{k\times p}, \tag{13}$$

where $\mathbf{N}_{n\times k} = \mathbf{Z}_{n\times p}\ \mathbf{E}_{p\times k}\ \mathbf{M}_{k\times k}$ is the mixing matrix corresponding to the full-dimensional anomaly data. Figure 10 shows six of the k resulting independent basis images $\mathbf{C}_{k\times p}$.

Figure 11 shows the six projections of the anomaly data $\mathbf{Z}_{n\times p}$ onto the independent components in Figure 10 as six columns of the matrix $\mathbf{Z}_{n\times p}\mathbf{C}^{'}_{p\times k}$, where $\mathbf{C}^{'}_{p\times k}$ denotes the transpose of $\mathbf{C}_{k\times p}$. The resulting time series are the analogues of the state space PC time series in Figure 5.

Figure 12 presents the time series from the last panel in Figure 11 along with the Niño 3.4 index. The correlation between the two series is 0.5788, slightly higher than the corresponding value between the best PC time series and the Niño 3.4 index. Shifting the IC time series back a few months increases the value of the correlation coefficient as follows: corr(1)=0.6895, corr(2)=0.7410, corr(3)=0.7654, corr(4)=0.7587, then decreases monotonically further. Figure 13 presents the IC time series shifted back by 3 months, along with the Niño region 3.4 index.

In addition to applying ICA to the first k eigenvectors, we also experimented with an alternative ICA model based on the first k state-space principal components in Eq. (3). This implementation, suggested in [1], leads to a decomposition of the form

$$\hat{\mathbf{Z}}^{(k)}_{n\times p} = \mathbf{D}_{n\times k}\mathbf{B}_{k\times p}, \tag{14}$$

where the independent components are the coefficients $\mathbf{D}_{n\times k}$ that linearly combine the basis images $\mathbf{B}_{k\times p}$. The resulting basis images showed similar characteristics to the basis images obtained by PCA. The highest correlation value

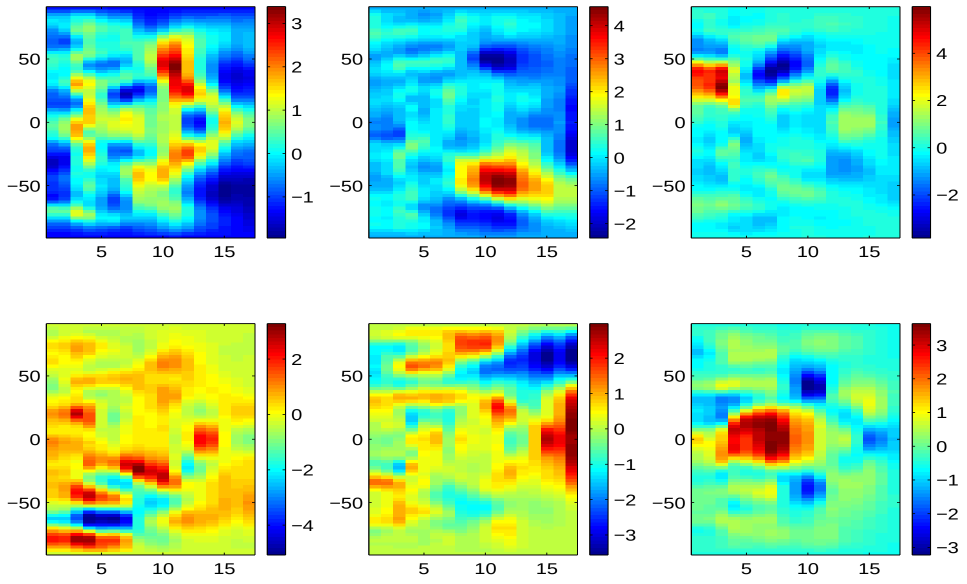


Figure 10: Six independent component bases obtained from the first k=22 eigenvectors of the anomaly data scatter matrix.

between the zonal anomaly data projected onto the bases in Eq. (14) and the Niño 3.4 index was 0.6031, achieved at one of the series lagged by five months. The previous two methods, PCA and ICA after dimension reduction with PCA, resulted in slightly higher correlations.

5 Summary

In this paper, we described two different techniques to reduce the dimension of climate data sets. The first few principal components in PCA provide a compact representation of the data, without significant loss of information. In addition, they also identify the main modes of variation in the data. The first principal component accounts for 20% of the total variation, and it represents an ENSO component. We also reported results with the more novel ICA method. Using ICA alone on the data proved to be problematic, but our initial results on combining ICA with PCA are promising. The simple ICA model used in combination with PCA for dimension reduction resulted in a superior ENSO signal estimate than PCA alone.

Our future work includes interpreting the remaining PC and IC estimates in terms of atmospheric processes; determining the sensitivity of the results to k, the number of components to estimate; incorporating constraints on the shapes of the sought components (for example, large volcano eruptions can only

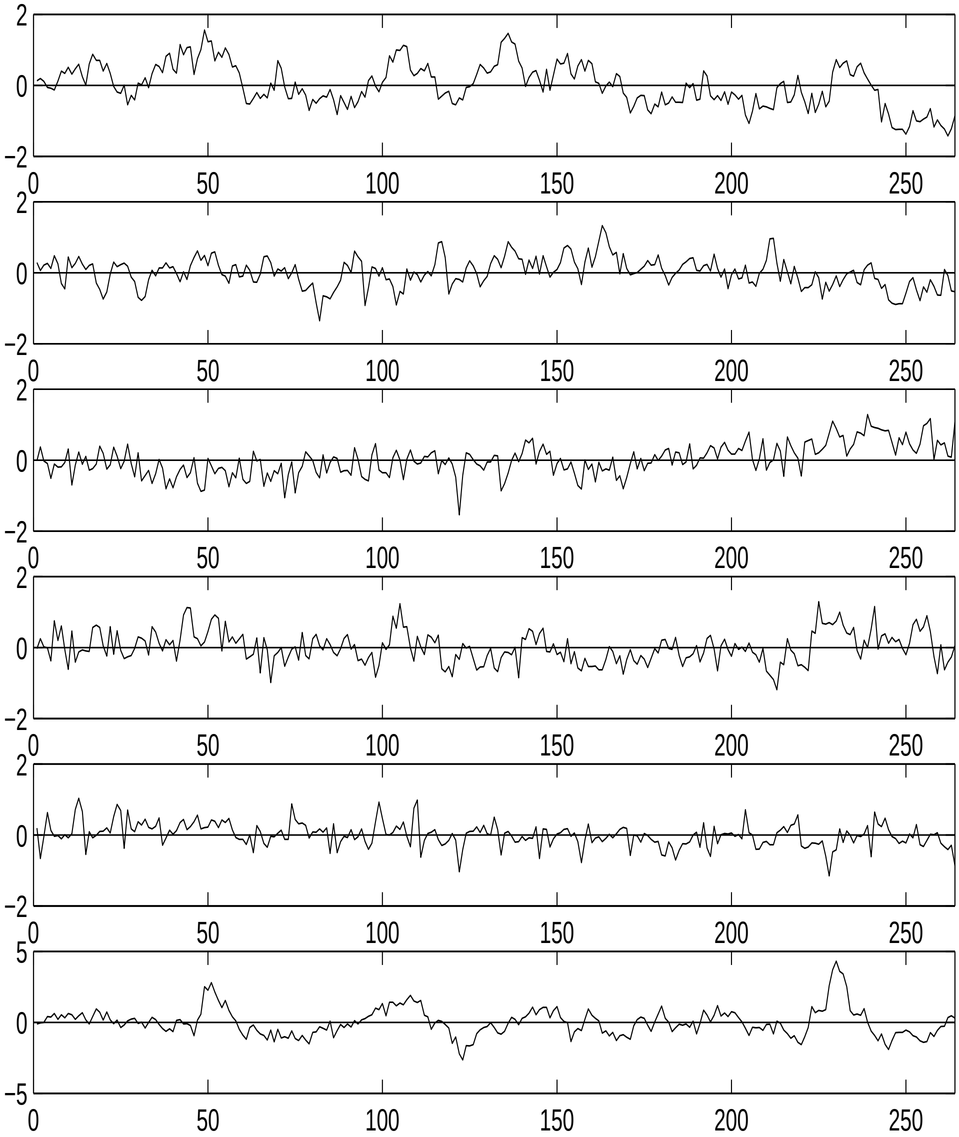


Figure 11: 6 state space IC time series.

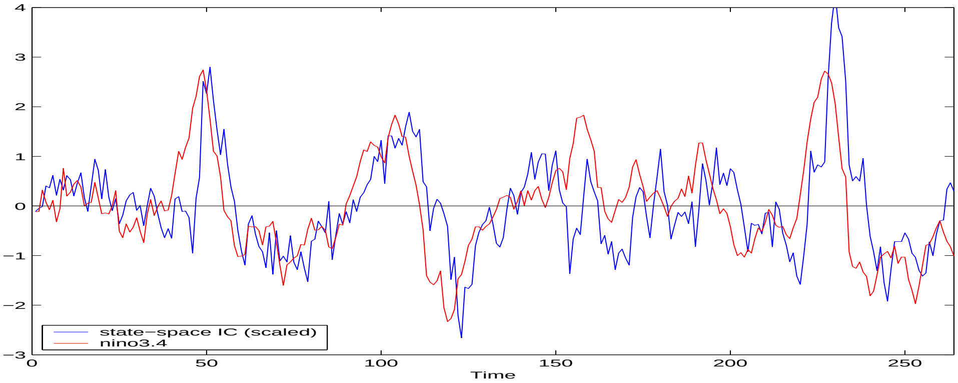


Figure 12: State space IC time series (scaled), and Niño region 3.4 index. Correlation coefficient is 0.58.

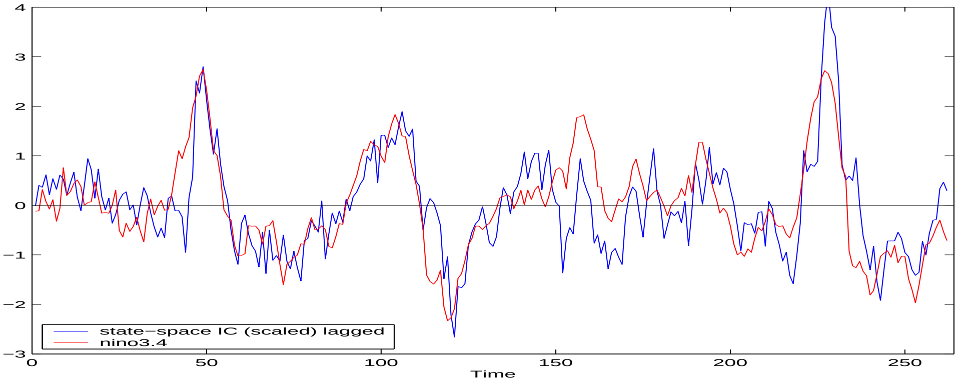


Figure 13: State space IC time series (scaled) with a three month lag, and Niño region 3.4 index. Correlation coefficient is 0.76.

decrease the temperature, but not increase it); and investigating possible nonlinearities in the mixing process.

Acknowledgments

We thank Dr. Benjamin D. Santer from the Program for Climate Model Diagnostics and Intercomparison at Lawrence Livermore National Laboratory for introducing us to the problem, for providing the climate data, and for his numerous insights and suggestions.

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, UCRL-JC-151613. This work was partially funded by the SciDAC program in the DOE Office of Advanced Scientific Computing Research.

References

[1] M. S. Bartlett, H.M. Lades, and T.J. Sejnowski. Independent component representations for face recognition. In *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging III*, volume 3299, pages 528–539, 1998.

[2] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, 2001.

[3] R. Kistler, E. Kalnay, et al. The NCEP/NCAR 50-year reanalysis. *Bull. Am. Met. Soc.*, 82:247–267, 2000.

[4] R.W. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography.* Number 17 in Developments in Atmospheric Science. Elsevier Science, Amsterdam, 1988.

[5] B.D. Santer et al. Accounting for the effects of volcanoes and ENSO in comparisons of modeled and observed temperature trends. *Journal of Geophysical Research*, 106(D22):28,033–28,059, November 27 2001.